

# DETECTING LOCAL SEMANTIC CONCEPTS IN ENVIRONMENTAL SOUNDS USING MARKOV MODEL BASED CLUSTERING

*Keansub Lee, Daniel P. W. Ellis\**

LabROSA, Dept. of Elec. Eng.  
Columbia University, USA  
{kslee, dpwe}@ee.columbia.edu

*Alexander C. Loui*

Kodak Research Laboratories  
Eastman Kodak Company, USA  
alexander.loui@kodak.com

## ABSTRACT

Detecting the time of occurrence of an acoustic event (for instance, a cheer) embedded in a longer soundtrack is useful and important for applications such as search and retrieval in consumer video archives. We present a Markov-model based clustering algorithm able to identify and segment consistent sets of temporal frames into regions associated with different ground-truth labels, and simultaneously to exclude a set of uninformative frames shared in common from all clips. The labels are provided at the clip level, so this refinement of the time axis represents a variant of Multiple-Instance Learning (MIL). Evaluation shows that local concepts are effectively detected by this clustering technique based on coarse-scale labels, and that detection performance is significantly better than existing algorithms for classifying real-world consumer recordings.

*Index Terms*— Environmental Audio, Audio Segmentation, Markov Models, Multiple Instance Learning

## 1. INTRODUCTION

Short consumer videos – casual recordings of people’s daily lives – are being created in huge numbers with today’s pocket cameras, and are extensively available on web sites like YouTube. These video clips contain a great deal of rich information relating to locations, activities, occasions, objects, etc., and consequently present many new opportunities for automatic extraction of semantic concepts to be used in intelligent browsing and retrieval systems. These concepts have diverse characteristics in terms of consistency, frequency and interrelationships. For example, concepts such as “music” or “crowd” typically persist over a large proportion of any clip to which they apply, and hence should be well represented in the global feature patterns, such as the mean and covariance of per-frame features of a clip. However, a “cheer” appears as a relatively small segment within a clip (at most a few seconds) which means that the global statistics of a longer clip

\*This work was supported by the NSF (grant IIS-0238301), the Eastman Kodak company, and EU project AMIDA (via the International Computer Science Institute).

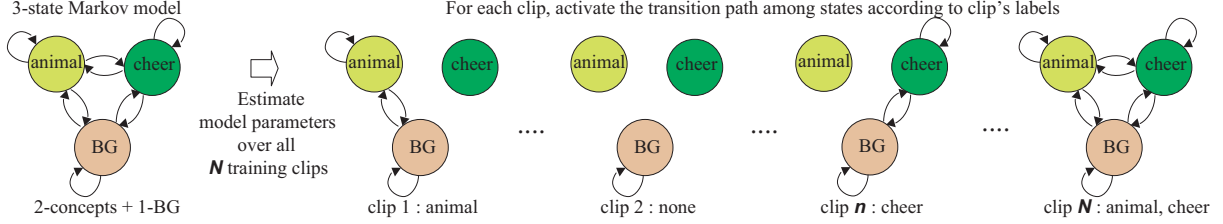
may fail to distinguish it from others. This paper addresses the problem of detecting such local patterns embedded in a global background soundtrack. In particular, we examine the case where we have training labels to indicate when examples of the concepts are present, but these labels are available only at the clip level (such as tags applied to YouTube videos), and therefore do not provide any more detailed information on the timing of the local events within the clip.

Multiple instance learning (MIL) has been successfully used to learn robust models from this kind of weak annotation across different levels of granularity. In MIL, each bag (e.g. an entire image or soundtrack) is a collection of instances (e.g. local feature vectors). Annotation is given at the bag level reflecting the label of one or more instances in that bag. If at least one instance is positive, the corresponding bag is labeled as positive. On the other hand, a bag is tagged as negative only when all instances in the bag are negative. The goal is to learn a set of instance points that are close to positive bags and simultaneously far away from negative bags. MIL, originally developed for applications in drug discovery [1], has been applied to content-based image retrieval, classification, and object detection [2, 3, 4].

In the next section, we describe a novel MIL approach, a Markov model-based clustering algorithm able to segment a set of temporal frames into multiple local regions (concepts) associated with different ground-truth labels tagged at the clip level, and simultaneously to exclude uninformative “background” frames shared in common from all clips. Evaluation and conclusions are presented in section 3 and 4 respectively.

## 2. LOCAL CONCEPT DETECTION

Our system starts with a basic frame-level feature, Mel-frequency Cepstral Coefficients (MFCCs), that are commonly used in speech recognition and other acoustic classification tasks. The single-channel (mono) soundtrack of a video is first resampled to 8kHz, and then a short-time Fourier magnitude spectrum is calculated over 25ms windows every 10ms. The spectrum of each window is warped to the Mel frequency scale, and the log of these auditory spectra is decorrelated



**Fig. 1.** Clustering temporal frames into concept-related segments by modifying the Markov model transition matrix.

into MFCCs via a discrete cosine transform. After the initial MFCC analysis, each video’s soundtrack is represented as a sequence of 21 dimensional MFCC feature vectors.

We then train a hidden Markov model (HMM) with Gaussian mixture emission models to learn the concepts. Each concept is a distinct state in the model, and in addition one or more “global background” states are included. The assumption here is that each feature vector can be associated with a particular concept (state), but through the time sequence of features in an entire clip, multiple different concepts may be expressed at different times. The model is learned via conventional Baum-Welch Expectation Maximization (EM), but for each clip the transition matrix is modified to ensure that only the states for the concepts specified in the labeling of that video and the global background states will be updated; transitions to all other concept states are set to zero. Figure 1 illustrates this idea for a 3-state Markov model.

The training process maximizes the likelihood of all frames using only the states allowed by the relevant clip-level annotations (and the global background states). It should result in states being used to model frames that are most relevant to those labels, with less informative frames being absorbed by the background models. Thus, the procedure achieves both clustering of frames that relate to each concept label, and produces a model that can be used to discriminate relevant sounds from uninformative background.

### 2.1. Markov Model-based Clustering

The HMM is parameterized by a set of parameters,  $\theta = \{\pi, A, \phi\}$  where  $\pi$ ,  $A$  and  $\phi$  indicate the prior, transition and emission probabilities of states. We begin by considering a single clip  $n$ . Assume that  $C_n$  denotes a  $K$ -dimensional annotation vector for a clip  $n$  in which each component,  $C_n(k) \in \{0, 1\}$  for  $k = 1, \dots, K$ , indicates the presence or absence of the  $k^{\text{th}}$  concept tagged by a human, and the  $K$  is the total number of concepts. Each concept can be present or absent independently in a clip. In our system, we annotated each training clip with 25 concepts as described in the section 3. We add 1, 2, or 4 states for the global background whose labels are set to be true (1) for all training clips; adding more background states allows for greater variety for this category, which we expect to account for the majority of the data. Thus,  $K$  is 26, 27 or 29.

The  $K \times K$ - dimensional transition matrix  $A$  is controlled by the ground truth annotation  $C_n$  of the clip  $n$  to be able to selectively train only the parameters of states whose corresponding concepts appear in  $C_n$ . The transition matrix  $A_n$  of the clip  $n$  is modified from the original  $A$  so that:

$$A_n(i, j) = A(i, j), \text{ iff } C_n(i) \text{ and } C_n(j) == 1. \quad (1)$$

All other values are set to zero.  $A_n(i, j)$  is then normalized by rows to satisfy  $\sum_{j=1}^K A_n(i, j) = 1$ .

The remaining process is to estimate the parameters,  $\theta = \{\pi, A, \phi\}$ , using the EM (Expectation Maximization) algorithm for all  $N$  training clips. Assume that  $X_n$  denotes the observations for the clip  $n$  comprising a set of MFCC feature vectors  $\{x_{nt}\}$  for  $t = 1, \dots, T_n$ , where  $T_n$  is the total number of frames in the clip  $n$  and depends on the duration of the original video. For every clip, we apply the forward-backward algorithm on  $X_n$ , with the corresponding modified transition matrix  $A_n$ , to evaluate the marginal posterior distribution  $\gamma(z_{ntk})$  of the latent variable  $z_{ntk}$  which indicates that frame  $t$  of clip  $n$  was emitted by state  $k$ . We also estimate the joint posterior distribution  $\xi(z_{n,t-1,k}, z_{ntk})$  of two successive latent variables in the E-step. The parameters  $\theta = \{\pi, A, \phi\}$  are updated for the M-step:

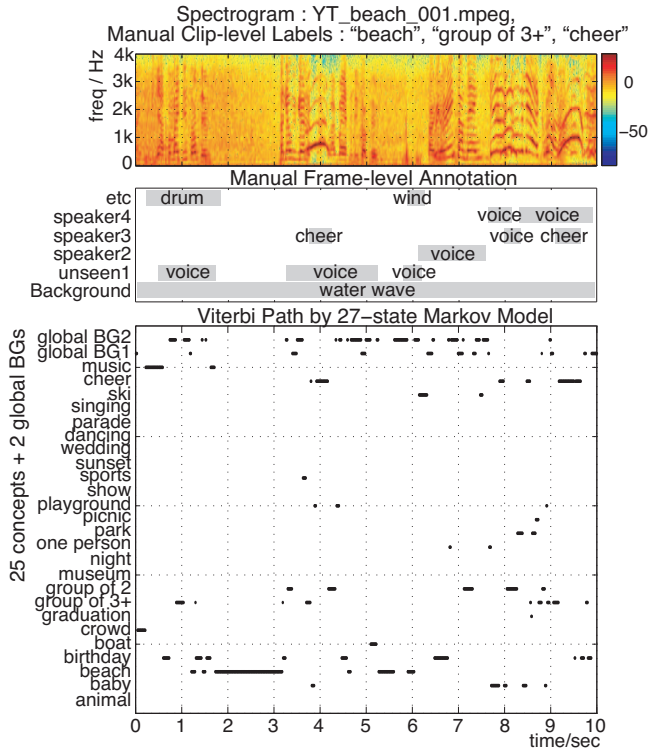
$$\pi_k = \frac{\sum_{n=1}^N \gamma(z_{n1k})}{\sum_{l=1}^K \sum_{n=1}^N \gamma(z_{n1l})} \quad (2)$$

$$A_{jk} = \frac{\sum_{n=1}^N \sum_{t=2}^{T_n} \xi(z_{n,t-1,j}, z_{ntk})}{\sum_{l=1}^K \sum_{n=1}^N \sum_{t=2}^{T_n} \xi(z_{n,t-1,j}, z_{ntl})} \quad (3)$$

The  $k^{\text{th}}$ ’s state emission probability,  $p(X; \phi_k)$ , is modeled by an  $M$ -component Gaussian mixture model (GMM):

$$p(X; \phi_k) = \sum_{m=1}^M w_{km} N(X | \mu_{km}, \Sigma_{km}) \quad (4)$$

Per-component weights  $w_{km}$ , means  $\mu_{km}$ , and covariances  $\Sigma_{km}$  are also updated using  $\gamma(z_{ntk})$ . We use  $M = 16$  for each state. The GMMs are initialized with a set of MFCC frames randomly selected from clips labeled with the appropriate concept. After learning the Markov model given the clip-level labels, the Viterbi algorithm is used to find the most probable sequence of states for a given sequence of MFCC frame in each testing clip as shown in Figure 2.



**Fig. 2.** Analysis of a soundtrack of a conversation at the beach. After Viterbi decoding, each frame is assigned to one of 27 concepts (25 primary plus two background states). Owing to the limitations of visual-based annotation, the speech from unseen speakers (e.g. the camera operator) is often not explicitly labeled, and so voice tends to fall into the global background as a sound common to all clips. However cheers and background beach noises are correctly identified.

### 3. EVALUATIONS

We tested our Markov clustering algorithm on the soundtracks of 1,873 videos downloaded from YouTube by using keywords (queries) relevant to the definition of our 25 concepts chosen to maximize usefulness in the final application, viability of obtaining manual labels, and viability of developing automatic recognizers. The concepts are organized into several broad classes including: activities (e.g. skiing, dancing), occasions (e.g. birthday, graduation), locations (e.g. beach, park), or particular objects in the scene (e.g. animal, baby, boat). Most concepts are intrinsically visual, although some, such as music and cheering, are primarily acoustic. The downloaded videos are filtered to retain only unedited, raw consumer videos whose averaged duration is 145 s. To ensure accurate labels, we manually reviewed every video and tagged it with the concepts that it contained. The video collections and labels are described in [5] in detail and available at <http://labrosa.ee.columbia.edu/projects/consumervideo/>.

To evaluate frame level performance, we further annotated the soundtracks of four object-related concepts (animal, baby, boat and cheer) to indicate the precise time segments that contain the sounds of those objects. The overall frame-level performance on this test data is presented in table 1 in terms of the frame-level accuracy,  $d'$  and Average Precision (AP). The accuracy rate is the proportion of 10 ms frames correctly labeled;  $d'$  is a threshold-independent measure of the separation between the two classes (presence and absence of the label) when mapped to two unit-variance Gaussian distributions, and AP is the Average of Precisions calculated separately for each true frame. Note that accuracy figures are high since in most cases there is a strong prior probability that any frame is negative (no relevant sound), so even labeling all frames negative would achieve high accuracy;  $d'$  and AP are less vulnerable to this bias.

Table 1 compares the frame-level concept classification results of semi-supervised MIL as described above with two supervised classifiers (SVM and Markov model) learned from frame-level annotations. Such detailed annotations are expensive to provide but provide an upper bound on what we hope to achieve with our less annotation-intensive MIL approach. The SVM classifier using an RBF kernel is trained on a set of manually annotated positive and negative frames for the corresponding concept, and then tested on other videos with using zero as threshold for deciding whether or not a concept is present. In the Markov model system, the GMMs are also initialized with a set of hand-labeled true frames of each concept. Severely biased priors between concept and non-concept frames, e.g. 2.44% for cheer, are known to be a challenge for SVM classifiers, so a lot of concept frames are wrongly lumped in with the background. The Markov model outperforms the SVM since it benefits from directly choosing among all the concepts at the same time.

For further comparison, we also report the frame-level performance of the ‘1G+KL with SVM’ system from [6], which trains an SVM classifier using a symmetrized Kullback-Leibler (KL) distance calculated on single, full-covariance Gaussian distributions fit to MFCC features over the entire clip. Here, to get a comparable sub-clip level time labeling, we divide the soundtrack into 1 s segments and classify each one. The resulting distance-to-boundary values from the SVM are shifted due to the change of segment’s length, so we try several different thresholds. The ‘‘T0’’ column in the table gives the results when classification is based on the standard SVM threshold of 0, which show the negative impact of this shift. Thus, we experiment with various other set of the threshold, shown in the subsequent columns: ‘‘T50’’ sets the threshold at the 50<sup>th</sup> percentile of the values within the clip, meaning that exactly half the labels in each test clip will be labeled positive. ‘‘T26S’’, ‘‘T27S’’, and ‘‘T29S’’ instead choose the percentile as the actual number of frames detected by the Markov clustering system with the corresponding number of states, as an upper-bound comparison.

**Table 1.** Frame-level concept classification performance on YouTube videos. The number below each concept indicates how many of the clips tagged with the concept actually contained relevant sounds; Values in columns 4 through 15 represent means of the frame-level performance over 5-fold cross-validated experiments: At each fold, the classifier is trained on 40% of the data, tuned on 20%, and then tested on the remaining 40% selected at random at the clip level.

Concept (# with sound)	Frame Prior (%)		Supervised (frame-level labels)				Semi-supervised (clip-level labels)							
			SVM w/ RBF	Markov Model			"1G+KL"+ SVM Classifier					Markov Clustering		
				26S	27S	29S	T0	T50	T26S	T27S	T29S	26S	27S	29S
animal (21/62 clips)	0.22	acc.(%)	74.8	98.1	98.5	<b>99.2</b>	36.6	72.9	98.5	98.6	99	98.4	98.5	98.9
		$d'$	0.24	0.29	0.38	0.12	0.15	0.47	0.38	0.31	0.3	0.57	<b>0.65</b>	0.3
		AP(%)	0.2	0.25	0.35	0.25	<b>0.67</b>					0.47	0.52	0.37
baby (43/112 clips)	0.4	acc.(%)	86	96.9	97.3	97.7	<b>99.6</b>	50.3	96.9	97.1	97.9	97	97.2	98
		$d'$	1.12	1.2	1.26	<b>1.3</b>	0	1.1	0.65	0.71	0.59	0.93	0.92	0.96
		AP(%)	<b>4.5</b>	3	2.7	3.5	0.73					1.64	1.83	1.65
boat (41/89 clips)	1.62	acc.(%)	92.7	97.3	97.7	<b>97.9</b>	58.8	50.7	96.5	97.1	97.6	97.1	97.6	<b>97.9</b>
		$d'$	0.88	1.47	1.34	1.24	0.47	0.61	0.17	0.13	0.2	1.3	<b>1.35</b>	1.3
		AP(%)	5.4	<b>12.2</b>	11.2	9.1	2.23					9.7	10.8	9
cheer (388/388 clips)	2.44	acc.(%)	46.8	94.8	95.2	95.4	<b>97.6</b>	52	92.7	93.2	93.5	95	95.3	95.4
		$d'$	1.62	<b>1.92</b>	1.91	<b>1.92</b>	0.15	1.38	0.1	0.11	0.11	1.77	1.76	1.72
		AP(%)	20.2	29.5	29.5	<b>29.8</b>	4.37					25.4	25.5	24.1

#### 4. DISCUSSION AND CONCLUSIONS

The results of the proposed Markov clustering based on clip-level annotation are given in the final three columns, for systems with 26, 27, or 29 states (i.e. 1, 2, or 4 background states). From the first column, we see that each label includes a significant number of clips labeled as relevant that do not in fact contain any relevant soundtrack frames (because the object makes no sound). Since the MIL approach is founded on the assumption that positive bags contain at least some positive examples, such incorrect annotations are a major factor degrading performance. We see that performance varies depending on the proportion of tagged clips that contain relevant audio frames (column 2). The animal concept, which has the worst result of  $d' = 0.65$ , contains relevant sound in only 34% (21/61) clips. Performance improves as the proportion of clips containing relevant sounds increases. Thus, "baby" with 43/112 = 38% relevant-sounding clips has  $d' = 0.92$ , "boat" (46%) has  $d' = 1.35$ , and "cheer" (100%) has  $d' = 1.76$ .

Another factor determining performance is the consistency of representative sounds for each concept. The "animal" concept covers many kinds of animal (e.g. dog, cat, fish etc.), and tends to have a very broad range of corresponding content. By comparison, "baby" is better because bay sounds (e.g. crying and laughing) are more specific than animal sounds. In the case of "boat", the relatively consistent engine noise is contained in a large proportion (46%) of relevant clips, leading to much better overall performance.

The best performance of the Markov clustering system occurs for cheering segments. We infer this is because the cheer concept is conveyed by acoustic information (leading to correct annotations), and its sound is consistent between different clips. The performance of Markov clustering with a random initialization ( $d' = 1.76$  and  $AP = 25.5\%$ ) is substantially better than a supervised SVM with RBF kernel ( $d' = 1.62$  and

$AP = 20.2\%$ ), and even is close to Markov modeling based on manual frame-level labels ( $d' = 1.92$  and  $AP = 29.8\%$ ), which we expect to be an upper bound. This provides the best illustration of the success of our Markov clustering in detecting local objects.

We have described detecting multiple local concepts from global annotations using Markov models. Experimental result shows that local concepts are effectively detected by this technique even based only on coarse clip-level labels, and that detection performance is significantly better than existing algorithms for real-world consumer recordings.

#### 5. REFERENCES

- [1] O. Maron and T. Lozano-Perez, "A framework for multiple instance learning," in *NIPS*, 1998.
- [2] Y. Chen and Z. Wang, "Image categorization by learning and reasoning with regions," in *Journal of Machine Learning Research*, 2004.
- [3] M. Naphade and J. Smith, "A generalized multiple instance learning algorithm for large scale modeling of multiple semantics," in *Proc. ICASSP*, Pittsburgh, PA, USA, Mar. 2005.
- [4] I. Ulusoy and C. M. Bishop, "Generative versus discriminative models for object recognition," in *Proc. CVPR*, San Diego, USA, 2005.
- [5] S.-F. Chang et al., "Kodak consumer video benchmark data set: concept definition and annotation," in *MIR workshop, ACM Multimedia*, Germany, Sep. 2007.
- [6] S.-F. Chang et al., "Large-scale multimodal semantic concept detection for consumer video," in *MIR workshop, ACM Multimedia*, Germany, Sep. 2007.